Systems Engineering Research Center

AI4SE & SE4AI Research and Application Workshop
September 17-18 2025, Washington DC

Two Quantitative Methods for Measuring and Comparing the Performance of Binary Classifiers

Mikel D. Petty, Ph.D.
Principal Research Scientist, Information Technology and Systems Center
Professor Emeritus, Computer Science
University of Alabama in Huntsville

Acknowledgements

- Sponsored by
  - U. S. Army Combat Capability Development Command Armaments Center
  - STTR A22B-T002, Phase II
    Verification, Validation, Assurance, and Trust of Machine Learning Models and Data for Safety-Critical Applications
- Managed by
  - Benjamin D. Werner, U. S. Army DEVCOM AC
  - Benjamin J. Schumeg, U. S. Army DEVCOM AC
- In collaboration with
  - OptTek Systems Inc., Boulder CO

Presentation outline

- Introduction
  - Overview
  - Classifiers and classifier confusion matrix
- Cost curves (Drummond and Holte, 2006)
  - Concept and definition
  - Examples
  - Comparing classifiers
- Safety scores (Salman et al., 2020)
  - Binary classifiers
  - Examples
  - Multiclass classifiers
- Backup (if requested and time allows)
  - Estimating safety score weights using MIL-STD-882E
  - Safety score weights and operating point parameters
  - Issues with safety scores as defined in source
  - Alternate Cost curve x and y formulas

Introduction

Overview

Task objective

- Find or develop and assess quantitative measures of classifier performance

Task components

- Implement, assess, and extend classifier Cost curves
- Implement, assess, and extend classifier Safety scores

Presentation content

- ~75% tutorial, explaining methods in the sources
- ~25% research, extensions of methods in the sources

Classifiers

Overview

- Classifier (AI model) presented with input instances
- Instances are each in one of two classes (binary), or each in one of k ($k \geq 2$) classes (multiclass)
- Classifier "predicts" or "classifies" class of each instance

Example binary classifier

- Target recognition
- Classes: target (positive), non-target (negative)

Target (positive)

Non-target (negative)

Classifier confusion matrix

		Predicted class: Positive	Predicted class: Negative
True class: Positive		Correct True positive *tp*	Type I error False negative *fn*
True class: Negative		Type II error False positive *fp*	Correct True negative *tn*

false negative rate	$FN = fn / (tp + fn)$
false positive rate	$FP = fp / (tn + fp)$
accuracy	$tp + tn / N$
precision	$tp / (tp + fp)$
recall	$tp / (tp + fn)$

Cost curves

Cost curves: Concept and definition (Drummond and Holte, 2006)

Operating point

- Numeric values represent classifiers' operational use conditions
- $p(+)$ = proportion of positive instances; $p(-) = 1 - p(+)$
- $C(-|+)$ = cost of misclassifying a positive instance as negative
- $C(+|-)$ = cost of misclassifying a negative instance as positive

Cost curves

- Graphical and quantitative measure to assess and compare binary classifiers
- Show classifiers' expected total cost during operation over full range of possible operational use conditions …
- … not just accuracy for single test with single input set
- x = function of operating point, represents operational use conditions
- y = function of x and classifiers' error rates
- Enables selecting best (lowest cost) classifier for anticipated operational use conditions

Cost curves: Examples, 1 of 5

Example classifiers and their rates

- N, P = trivial classifiers (N always negative, P always positive)
- A, B, …, G = notional non-trivial classifiers
- *FN*, *FP* = classifiers' false negative rate, false positive rate

Example 1

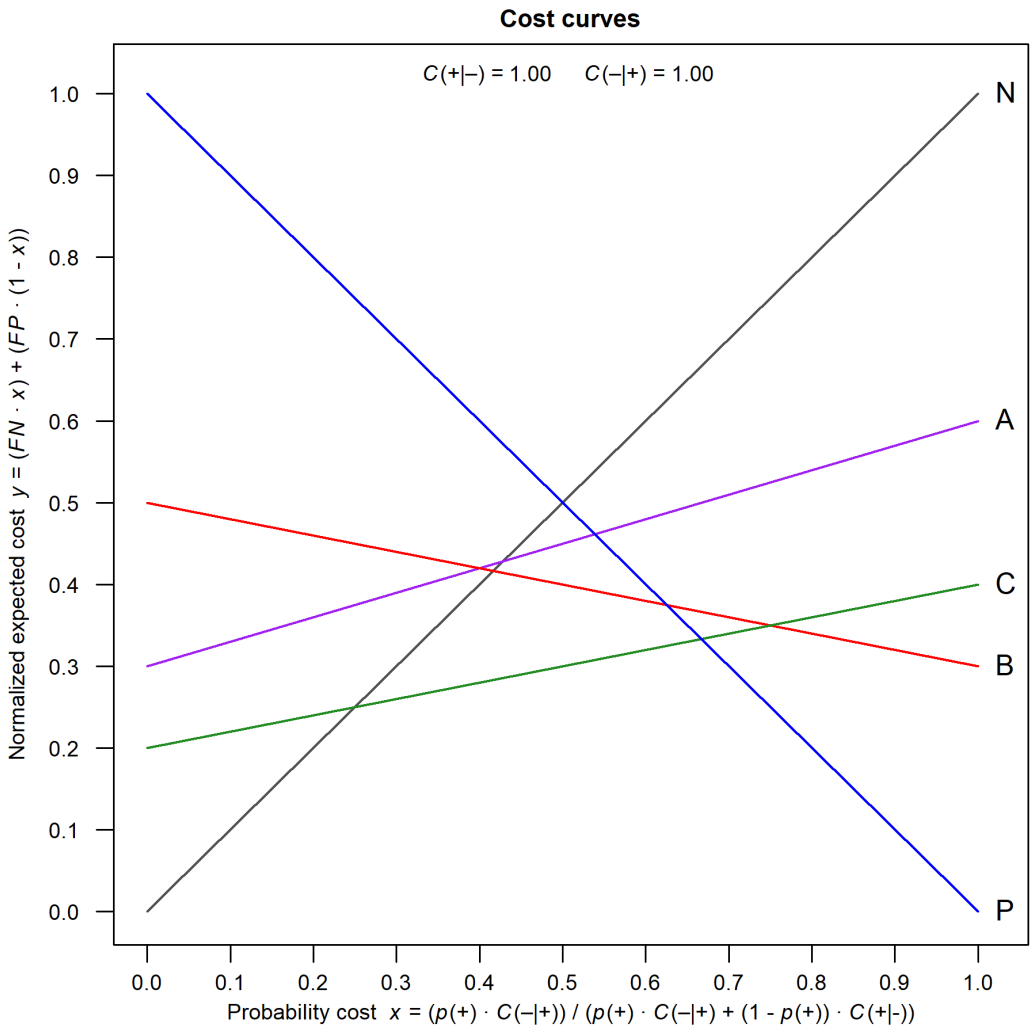
Classifier	*FN*	*FP*
N	1.0	0.0
A	0.6	0.3
B	0.3	0.5
C	0.4	0.2
P	0.0	1.0

Example 2

Classifier	*FN*	*FP*
N	1.00	0.00
D	0.84	0.05
E	0.60	0.15
F	0.30	0.35
G	0.15	0.50
P	0.00	1.00

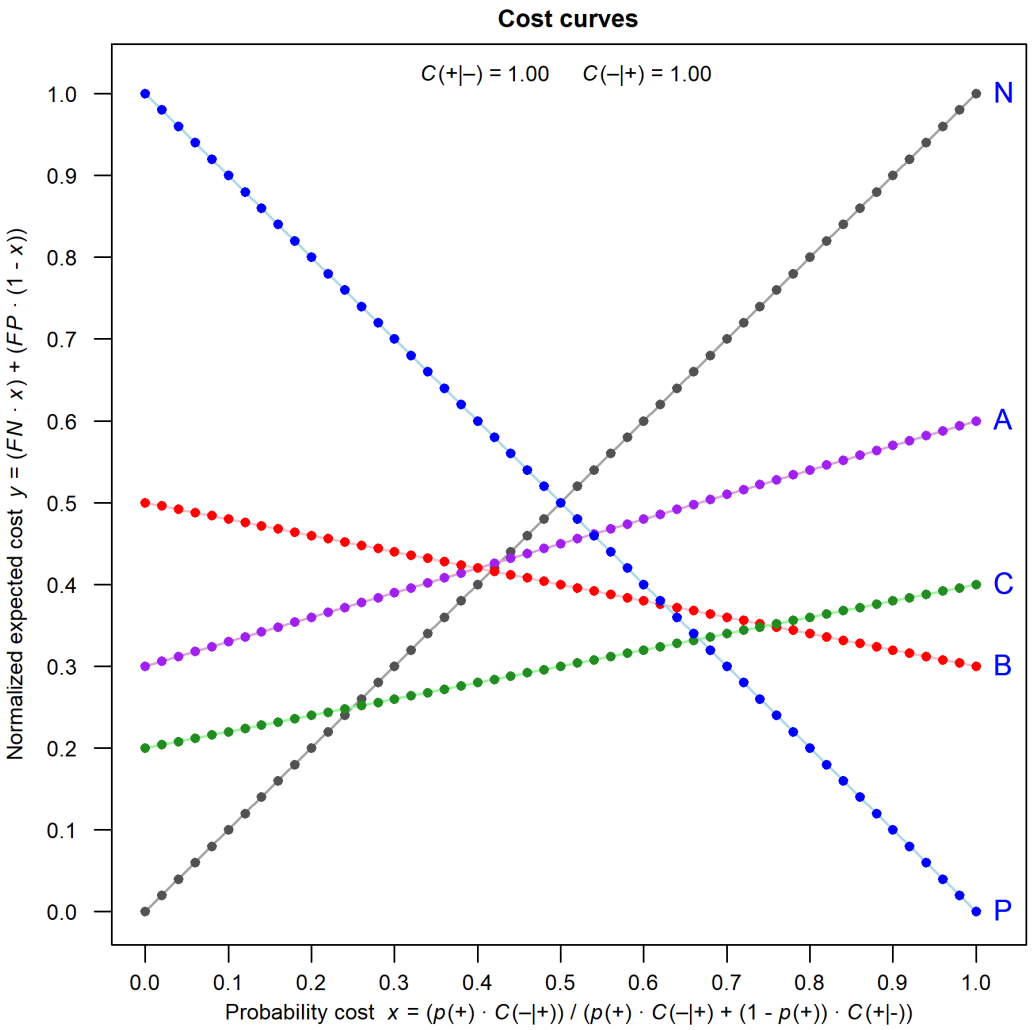
Cost curves: Examples, 2 of 5

Example 1



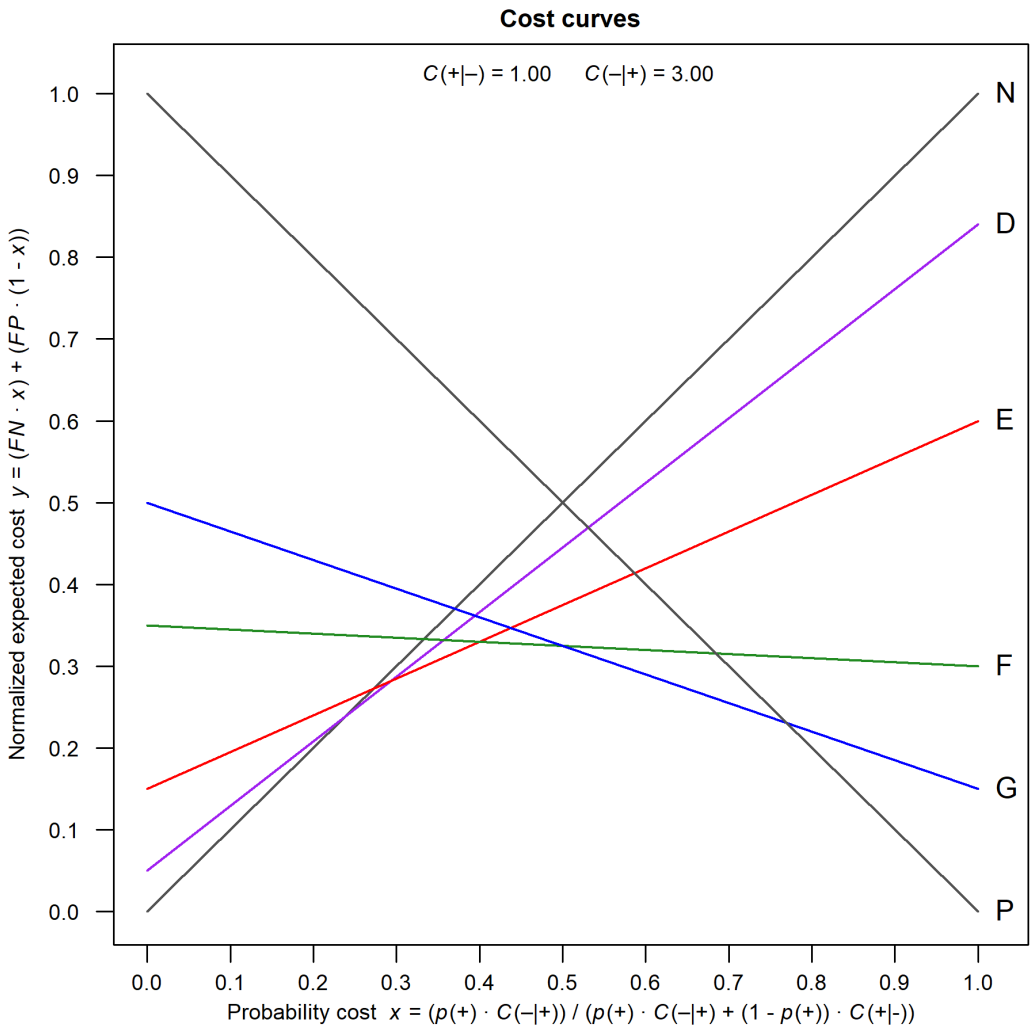
Cost curves: Examples, 3 of 5

Example 1



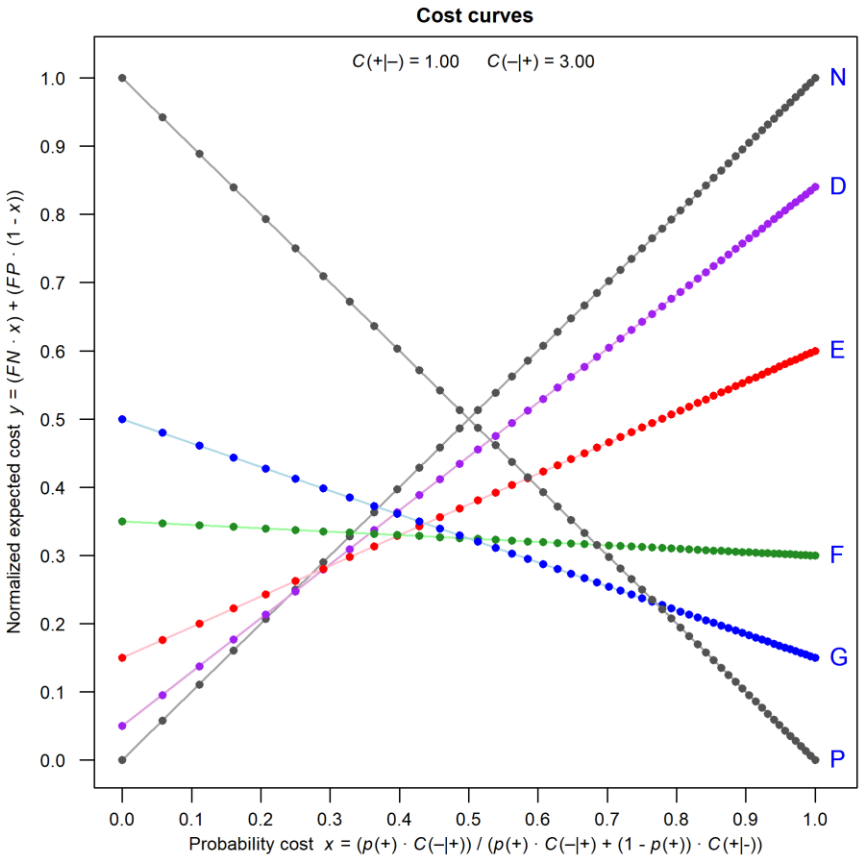
Cost curves: Examples, 4 of 5

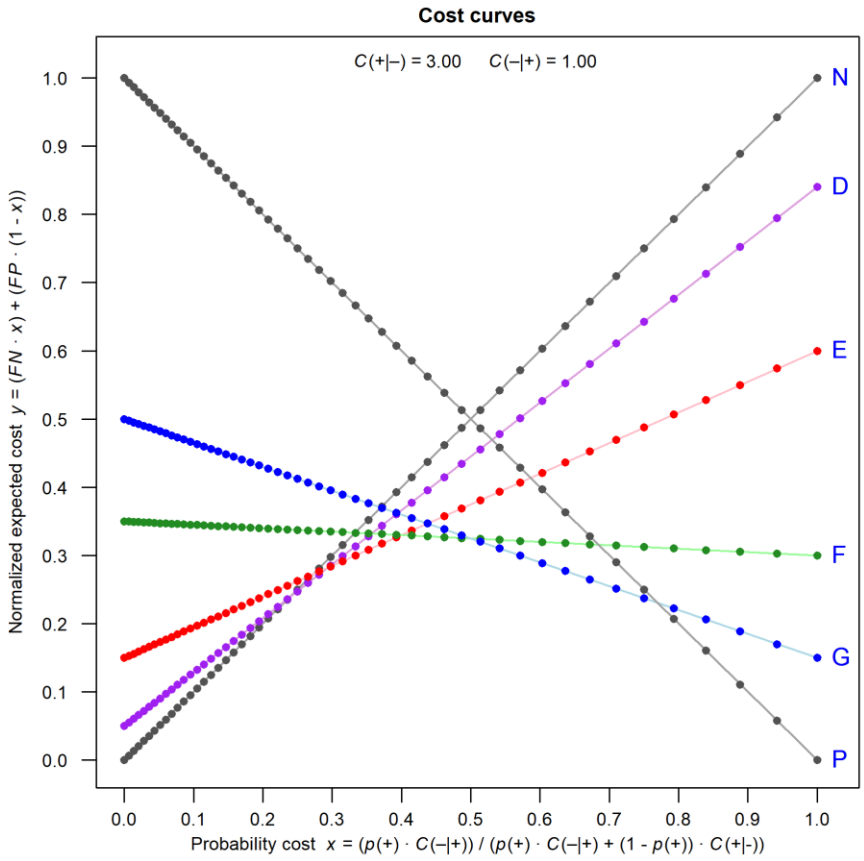
Example 2



Cost curves: Examples, 5 of 5

Example 2



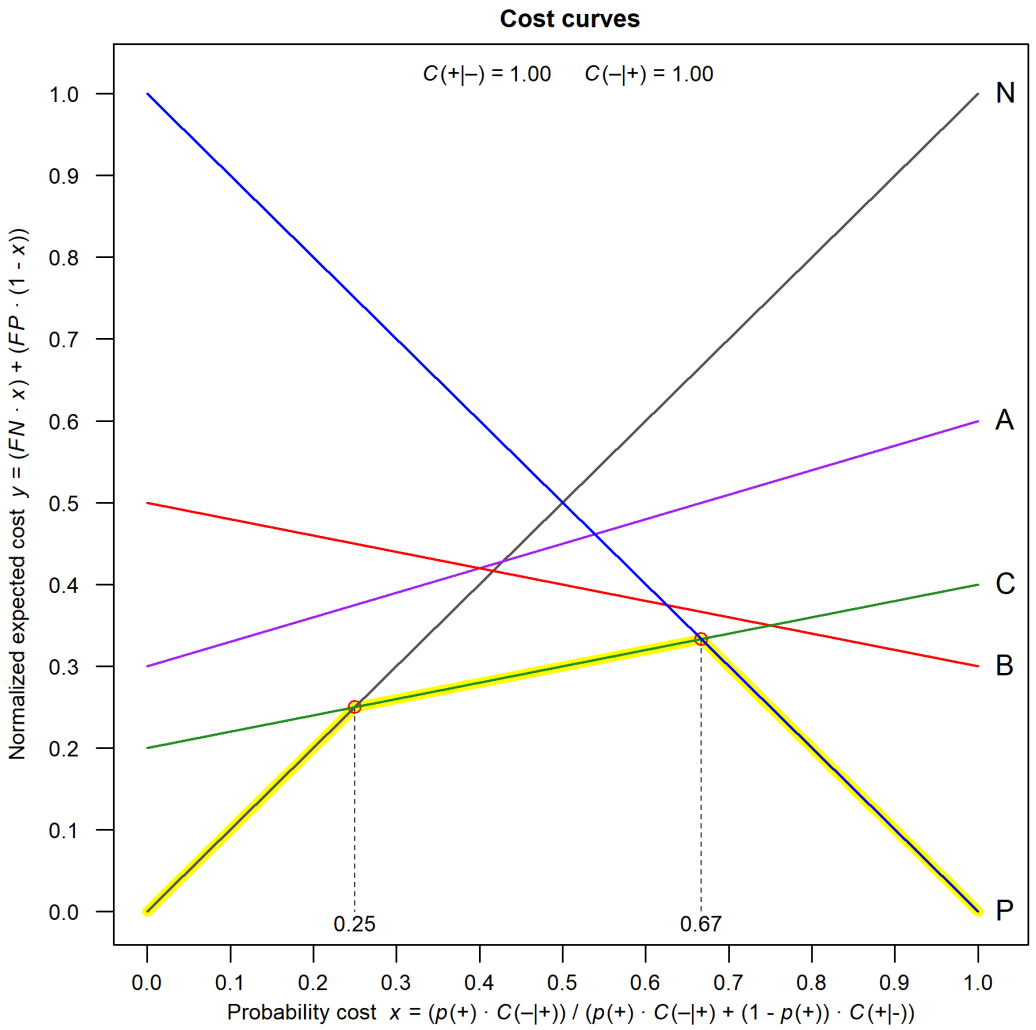


OptTek

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

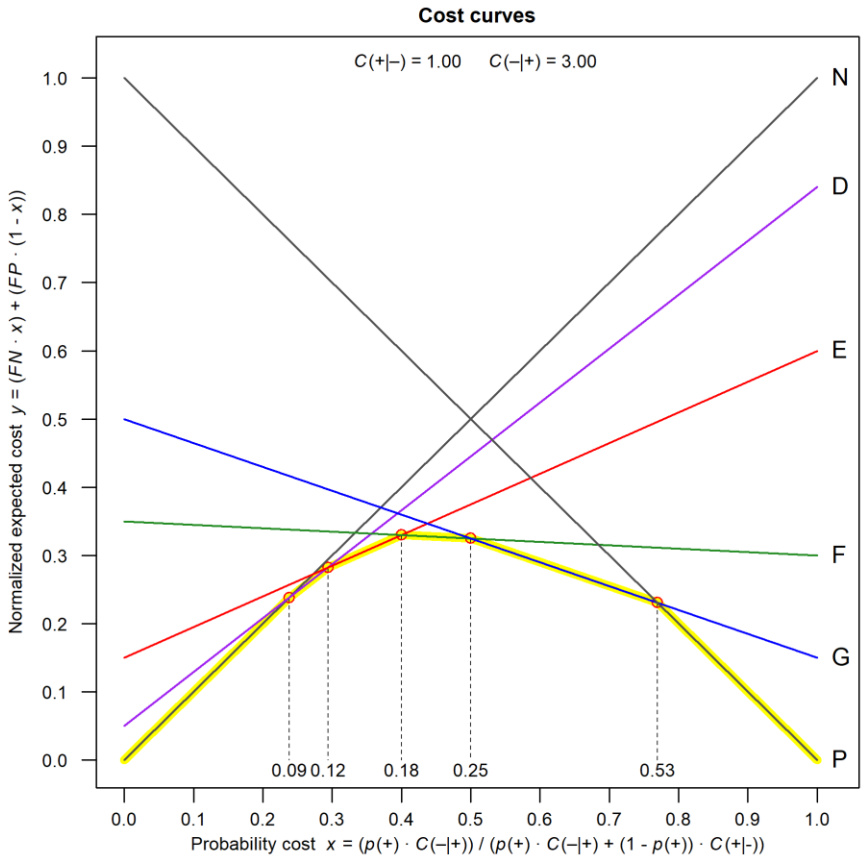
Cost curves: Comparing classifiers, 1 of 2

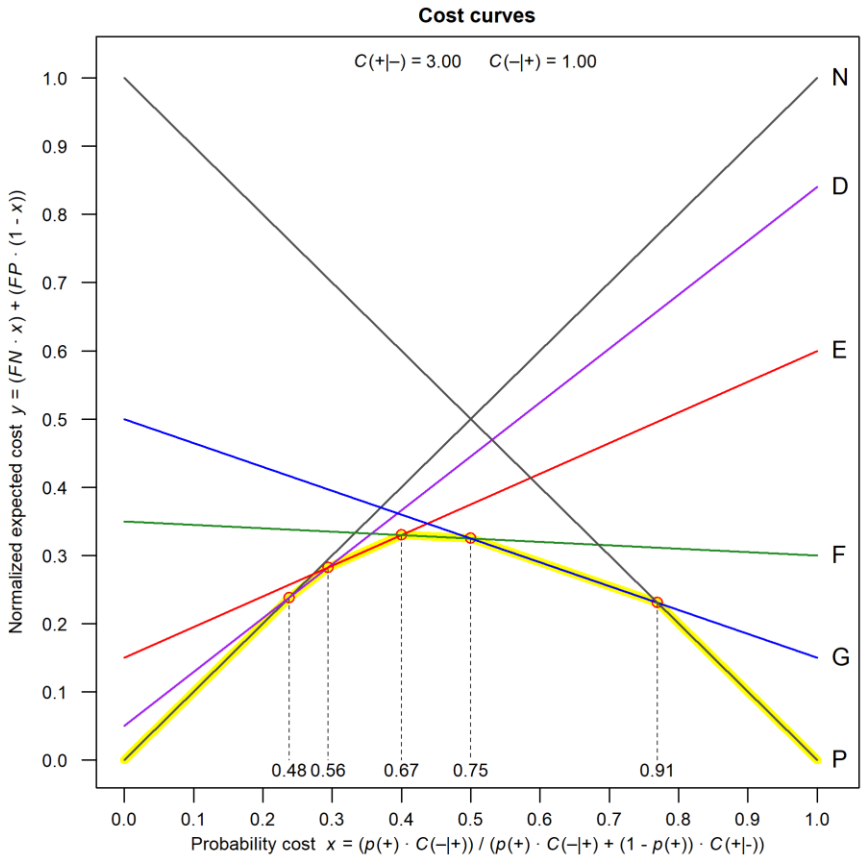
Example 1



Cost curves: Comparing classifiers, 2 of 2

Example 2





OptTek

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Safety scores

Safety scores: Binary classifiers, 1 of 4 (Salman et al., 2020)

Standard safety score formula

$$safety\ score = \frac{w_{tp}tp + w_{tn}tn}{w_{fp}fp + w_{fn}fn + w_{tp}tp + w_{tn}tn}$$

Values in formula

- Weights w_{tp}, w_{tn}, w_{fp}, w_{fn} represent cost of classifying an instance, both correct (w_{tp}, w_{tn}) and incorrect (w_{fp}, w_{fn}) classifications
- Costs estimated by SME based on operational use conditions, then normalized to get weights $\in [0, 1]$
- Counts tp, tn, fp, fn are counts of classifications; both correct (tp, tn) and incorrect (fp, fn) have costs in safety score
- Counts result from testing classifier

Issue with standard formula

- Counts results of testing with single dataset with fixed $p(+)$
- $\rightarrow$ standard safety score uninformative about other $p(+)$ values

Safety scores: Binary classifiers, 2 of 4

Enhanced safety score formula

$$A = w_{tp} \cdot (N \cdot p(+) \cdot (1 - FN)) + w_{tn} \cdot (N \cdot (1 - p(+)) \cdot (1 - FP))$$
$$B = w_{fp} \cdot (N \cdot (1 - p(+)) \cdot FP) + w_{fn} \cdot (N \cdot p(+) \cdot FN)$$
$$enhanced\ safety\ score = \frac{A}{A + B}$$

Estimated by SME based on expected operational use conditions

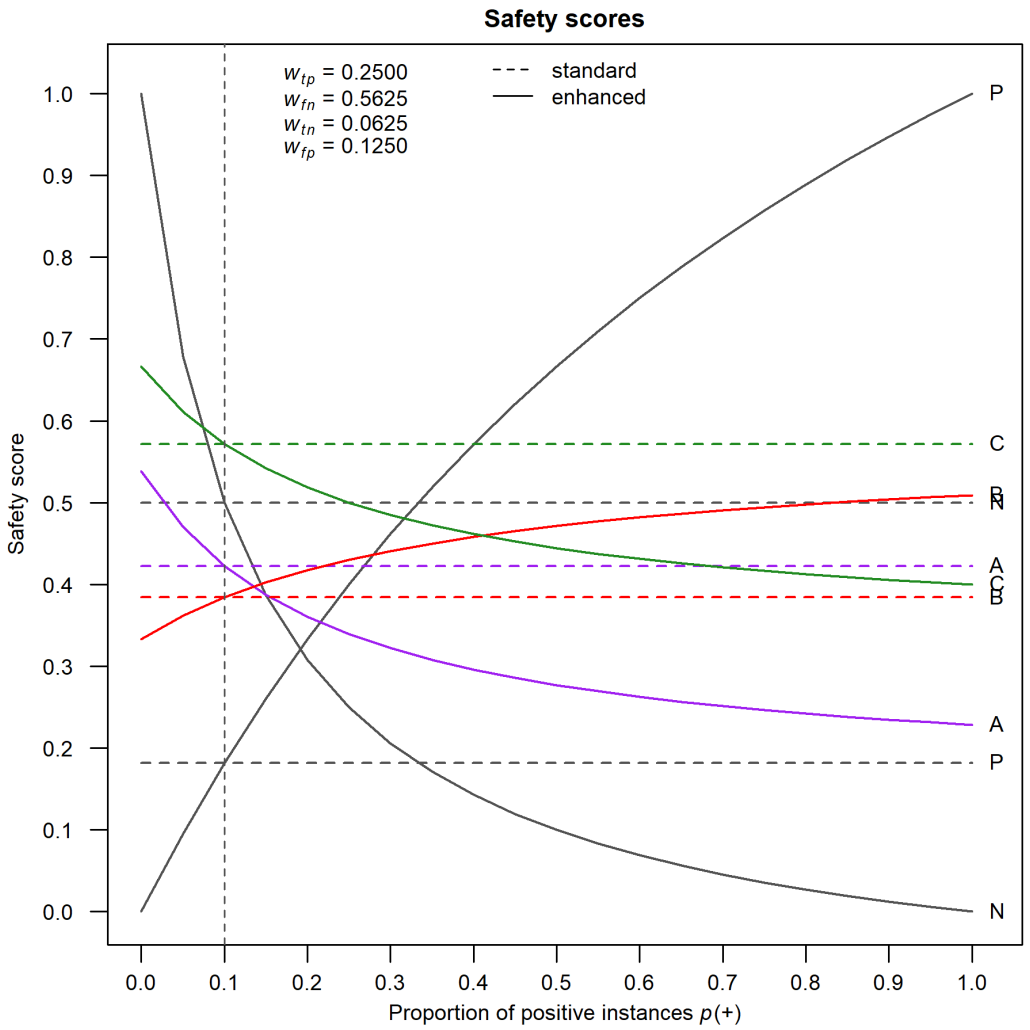
- Weights (classification costs) $w_{tp}, w_{tn}, w_{fp}, w_{fn}$
- Number of instances N
- Proportion of positive instances $p(+)$

Found by testing classifier

- False negative rate $FN = fn / (tp + fn)$
- False positive rate $FP = fp / (tn + fp)$

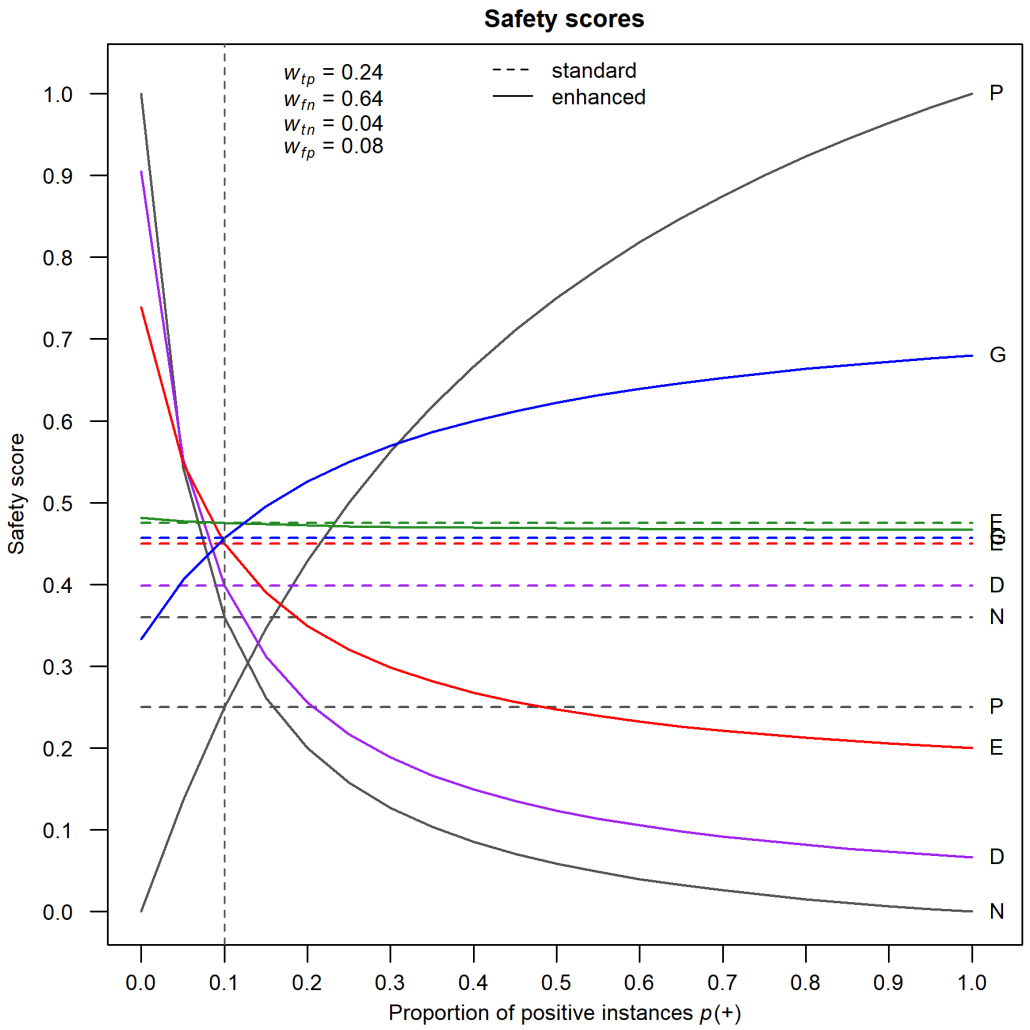
Safety scores: Binary classifiers, 3 of 4

Example 1



Safety scores: Binary classifiers, 4 of 4

Example 2



Safety scores: Multiclass classifiers, 1 of 4

Standard multiclass safety score formula

weights $W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & & \\ \vdots & & \ddots & \vdots \\ w_{k1} & & \cdots & w_{kk} \end{bmatrix}$ counts $C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & & \\ \vdots & & \ddots & \vdots \\ c_{k1} & & \cdots & c_{kk} \end{bmatrix}$

$$standard\ multiclass\ safety\ score = \frac{\sum_{i=1}^{k} w_{ii} \cdot c_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \cdot c_{ij}}$$

- k = number of classes, $k \geq 2$
- w_{ij} = weight (cost) of classifying an instance of class i as class j, estimated by SME
- c_{ij} = count of instances of class i classified as class j, results of testing classifier with single dataset

Safety scores: Multiclass classifiers, 2 of 4

Enhanced multiclass safety score formula

weights $W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & & \\ \vdots & & \ddots & \vdots \\ w_{k1} & & \cdots & w_{kk} \end{bmatrix}$ counts $C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & & \\ \vdots & & \ddots & \vdots \\ c_{k1} & & \cdots & c_{kk} \end{bmatrix}$

$$enhanced\ multiclass\ safety\ score = \frac{\sum_{i=1}^{k} w_{ii} \cdot c_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \cdot c_{ij}}$$

- k = number of classes, $k \geq 2$
- w_{ij} = weight (cost) of classifying an instance of class i as class j, estimated by SME
- c_{ij} = $N \cdot p(i) \cdot P_{ij}$ for $1 \leq i, j \leq k$, where
  - N = number of instances; estimated by SME
  - $p(i)$ = proportion of instances of class i; estimated by SME
  - P_{ij} = probability that instance of class i is classified as class j; found by testing classifier

Safety scores: Multiclass classifiers, 3 of 4

Example 1 (data)

$$\text{costs} = \begin{bmatrix} 1 & 2 & 4 & 8 \\ 2 & 1 & 2 & 4 \\ 8 & 4 & 0 & 2 \\ 16 & 8 & 2 & 0 \end{bmatrix} \quad W = \begin{bmatrix} 0.015625 & 0.03125 & 0.0625 & 0.125 \\ 0.03125 & 0.15625 & 0.03125 & 0.0625 \\ 0.125 & 0.0625 & 0 & 0.03125 \\ 0.25 & 0.125 & 0.03125 & 0 \end{bmatrix}$$

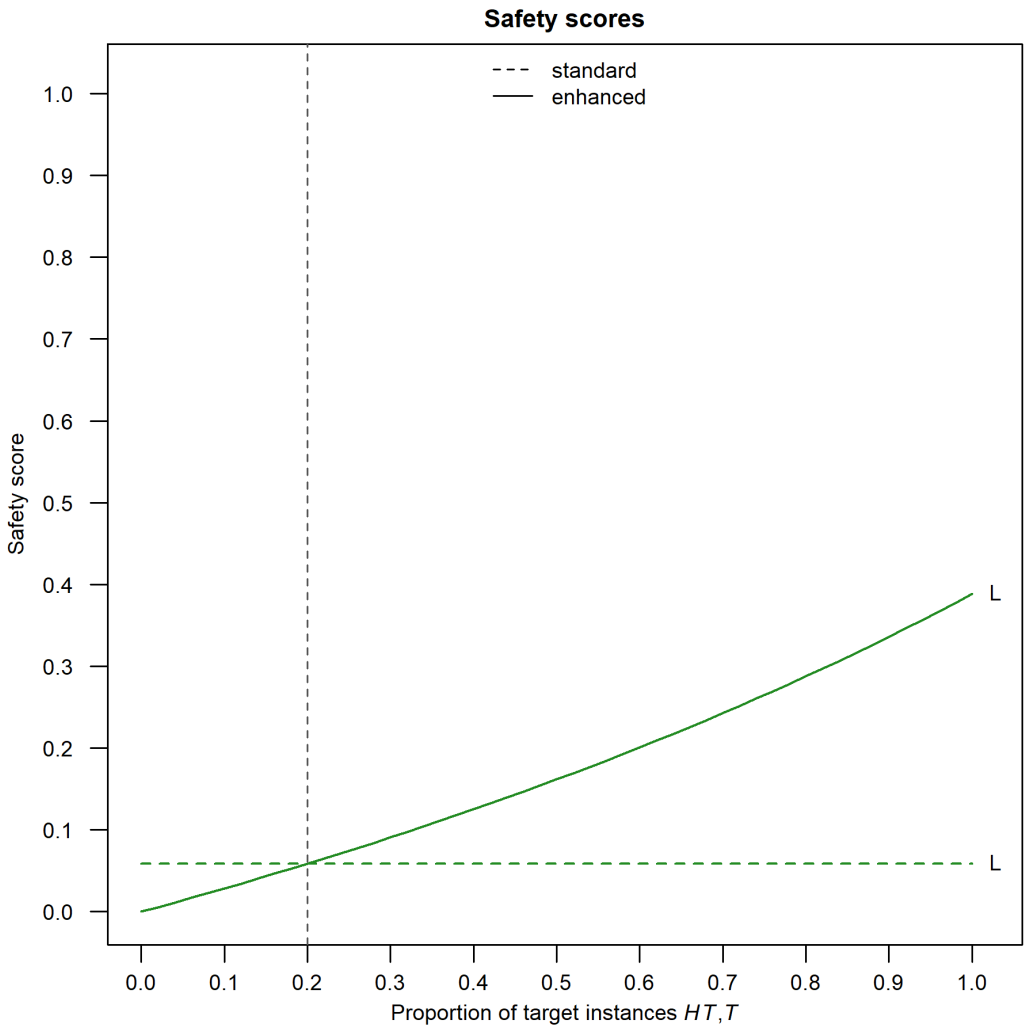
$$\text{probabilities } P = \begin{bmatrix} 0.65 & 0.20 & 0.10 & 0.05 \\ 0.20 & 0.60 & 0.15 & 0.05 \\ 0.10 & 0.15 & 0.50 & 0.25 \\ 0.10 & 0.15 & 0.30 & 0.45 \end{bmatrix}$$

proportions =

Class	Standard safety score		Enhanced safety score	
	Category proportion	Class proportion	Category proportion	Class proportion
High value target (HT)	0.2	0.2	p(target)	0.2
Target (T)		0.8		0.8
Non-target (NT)	0.8	0.8	$1 - p$(target)	0.8
High value non-target (HNT)		0.2		0.2

Safety scores: Multiclass classifiers, 4 of 4

Example 1 (plot)



Sources

- (Drummond and Holte, 2006)

  Drummond, C. and Holte, R. C. (2006), "Cost curves: An improved method for visualizing classifier performance", *Machine Learning*, Vol. 65, pp. 95-130, https://doi.org/10.1007/s10994-006-8199-5.

- (Salman et al., 2020)

  Salman, T., Ghubaish, A., Unal, D., and Jain, R. (2020), "Safety Score as an Evaluation Metric for Machine Learning Models of Security Applications", *IEEE Networking Letters*, Vol. 2, No. 4, December 2020, pp. 207-2011, https://doi.org/10.1109/LNET.2020.3016583.

Questions?

Backup

Estimating safety score weights using MIL-STD-882E

Description	Severity Category	Mishap Result Criteria
Catastrophic	1	Could result in one or more of the following: death, permanent total disability, irreversible significant environmental impact, or monetary loss equal to or exceeding $10M.
Critical	2	Could result in one or more of the following: permanent partial disability, injuries or occupational illness that may result in hospitalization of at least three personnel, reversible significant environmental impact, or monetary loss equal to or exceeding $1M but less than $10M.
Marginal	3	Could result in one or more of the following: injury or occupational illness resulting in one or more lost work day(s), reversible moderate environmental impact, or monetary loss equal to or exceeding $100K but less than $1M.
Negligible	4	Could result in one or more of the following: injury or occupational illness not resulting in a lost work day, minimal environmental impact, or monetary loss less than $100K.

Table I

Severity / Probability	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	High	High	Serious	Medium
Probable (B)	High	High	Serious	Medium
Occasional (C)	High	Serious	Medium	Low
Remote (D)	Serious	Medium	Medium	Low
Improbable (E)	Medium	Medium	Medium	Low
Eliminated (F)	Eliminated			

Table III

Description	Level	Specific Individual Item	Fleet or Inventory
Frequent	A	Likely to occur often in the life of an item.	Continuously experienced.
Probable	B	Will occur several times in the life of an item.	Will occur frequently.
Occasional	C	Likely to occur sometime in the life of an item.	Will occur several times.
Remote	D	Unlikely, but possible to occur in the life of an item.	Unlikely, but can reasonably be expected to occur.
Improbable	E	So unlikely, it can be assumed occurrence may not be experienced in the life of an item.	Unlikely to occur, but possible.
Eliminated	F	Incapable of occurrence. This level is used when potential hazards are identified and later eliminated.	Incapable of occurrence. This level is used when potential hazards are identified and later eliminated.

Table II

Severity description	Severity category	Dollar amount	Safety score weight
Catastrophic	1	50,000,000	0.90
Critical	2	5,000,000	0.09
Marginal	3	500,000	0.009
Negligible	4	50,000	0.001

Weights

Example mapping (notional)

$w_{fp} = 0.90$ (catastrophic) $\quad w_{fn} = 0.09$ (critical)

$w_{tp} = 0.009$ (marginal) $\quad w_{tn} = 0.001$ (negligible)

Safety score weights and operating point parameters

Safety score weight	Operating point parameter
w_{tp}	–
w_{tn}	–
w_{fp}	$C(+\mid -)$
w_{fn}	$C(-\mid +)$
–	$p(+)$

- $p(+)$ included in enhanced safety score formulas
- Including $w_{tp} = C(+\mid +)$ and $w_{tn} = C(-\mid -)$ in cost curves listed as "future work"

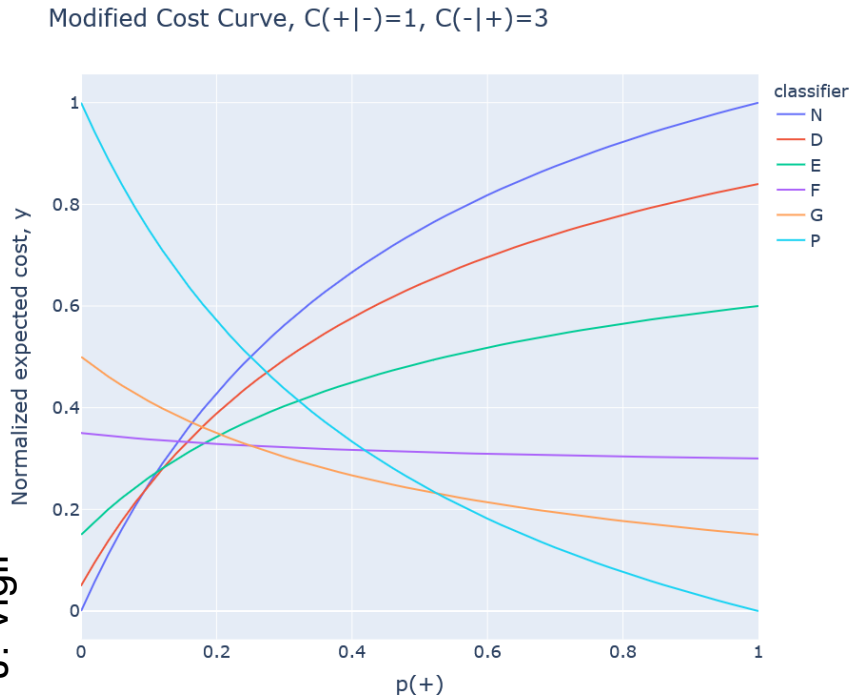
Issues with safety scores as defined in source

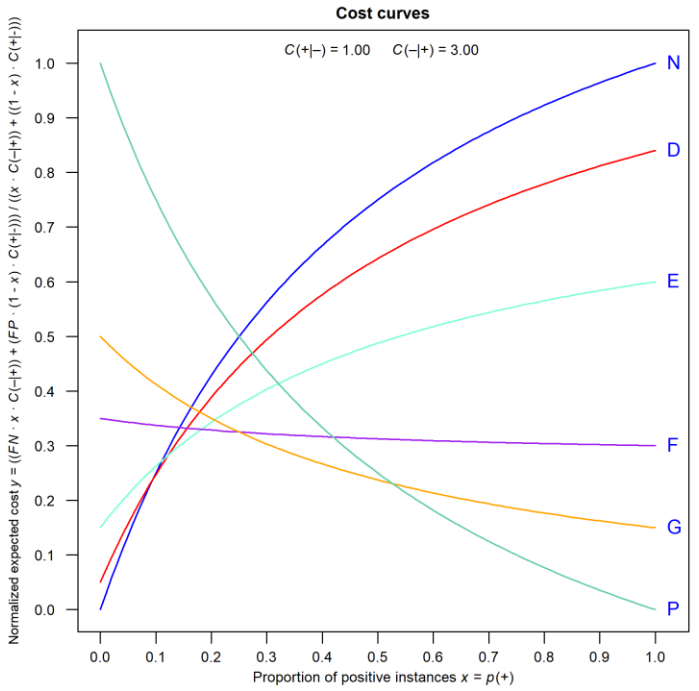
Safety score definition in (Salman et al., 2020)

- Formula does not include $p(+)$; addressed in enhanced safety score formulas
- Does not explain how to estimate weights; see MIL-STD-882E topic
- Assumes, without explicitly stating assumption, that all costs are positive
  - Costs $= 0 \rightarrow$ possible uninformative safety scores or divide-by-zero error
  - Costs $< 0 \rightarrow$ possible divide-by-zero error

Alternate Cost curve x and y formulas

J. Vigil





M. Petty

$$x = p(+)$$

$$y = \frac{(FN \cdot x \cdot C(-|+)) + (FP \cdot (1-x) \cdot C(+|-))}{(x \cdot C(-|+)) + ((1-x) \cdot C(+|-))}$$

OptTek

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

End